The Canadian Journal for the Scholarship of Teaching and Learning

Volume 8 | Issue 1 Article 9

March 2017

The Presence of Gender Disparity on the Force Concept Inventory in a Sample of Canadian Undergraduate Students

Magdalen Normandeau
University of New Brunswick, mnormand@unb.ca
Seshu Iyengar
University of New Brunswick, seshu.iyengar@unb.ca
Benedict Newling
University of New Brunswick, bnewling@unb.ca

https://doi.org/10.5206/cjsotl-rcacea.2017.1.9

The Presence of Gender Disparity on the Force Concept Inventory in a Sample of Canadian Undergraduate Students

Abstract
Concept inventories (CI) are validated, research-based, multiple-choice tests, which are widely used to assess the effectiveness of pedagogical practices in bringing about conceptual change. In order to be a useful diagnostic tool, a CI must reflect only the student understanding of the conceptual material. The Force Concept Inventory (FCI) is arguably the standard for testing conceptual understanding of Newtonian mechanics. Studies in the United States and United Kingdom have shown the existence of a gender gap in FCI scores and gains between male and female students. This study aimed to examine whether such a gap exists for Canadian students at a mid-sized university. Four-hundred and thirty-four men and 379 women taking first-term introductory physics courses from the past nine years were assessed with the FCI prior to and after receiving instruction. A gender gap in the pre-instruction and post instruction scores was revealed in favour of male students ($p < 0.01$). There also existed a gender disparity in the learning gains between the two tests, where males had significantly higher gains ($p < 0.01$), although the effect size was small. Further analysis found that both male and female students who studied in classes that included interactive engagement methods had somewhat higher gains than students in traditional lecture courses, but that the interactive engagement methods did not eliminate the gender gap between male and female students ($p < 0.01$). Our results sound a cross-disciplinary note of caution for anyone using concept inventories as research or self-assessment tools.

Les inventaires de concepts sont des questionnaires à choix multiples validés basés sur la recherche qui sont largement utilisés pour évaluer l'efficacité de pratiques pédagogiques en instaurant un changement conceptuel. Afin d'être des outils diagnostiques utiles, les inventaires de concepts doivent refléter uniquement la compréhension qu'a l'étudiant de la matière conceptuelle. Le « Force Concept Inventory (FCI) » est sans aucun doute la norme pour tester la compréhension conceptuelle de la mécanique newtonienne. Des études menées aux États-Unis et au Royaume-Uni ont montré l'existence d'un écart hommes-femmes dans les résultats du FCI et ainsi que dans les acquis. Cette étude vise à déterminer si un tel écart existe parmi les étudiants canadiens dans une université de taille moyenne. Un total de 434 hommes et 379 femmes inscrits à un premier cours d'introduction à la physique au fil des neuf dernières années ont été évalués avec le FCI au tout début et à la toute fin de la session. Les résultats ont révélé un écart hommes-femmes dans les résultats des tests, aussi bien ceux effectués avant le cours que ceux après le cours, en faveur des étudiants masculins ($p < 0.01$). Ils ont également révélé une disparité entre hommes et femmes dans les acquis d'apprentissage entre les deux tests : les hommes avaient atteint des acquis plus élevés ($p < 0.01$), bien que l'ampleur de l'effet ait été faible. Des analyses complémentaires ont montré que tant les hommes que les femmes qui avaient étudié dans des classes qui comprenaient des méthodes d'engagement interactif avaient obtenu davantage d'acquis que les étudiants qui avaient suivi des cours magistraux traditionnels, mais que les méthodes d'engagement interactif n'avaient pas éliminé l'écart hommes-femmes parmi les étudiants ($p < 0.01$). Nos résultats présentent une mise en garde à l'intention de ceux qui utilisent les inventaires de concepts en tant qu'outils de recherche ou d'auto-évaluation, quelle que soit leur discipline.

Keywords
physics education research, concept inventory, gender

Cover Page Footnote
The authors are grateful for the support of the University of New Brunswick (UNB) through the Teaching & Learning Priority Fund and the University Teaching Scholar program.

When our students walk into our classrooms, they come with years of experience during which they have observed and interacted with the world. From this, they have constructed their understanding of the world (see Bransford, Brown, & Cocking, 2000). While these personal mental models have their uses (e.g., "If I jump from higher up, I will be moving faster when I hit the ground, so I am more likely to hurt myself."), because they are based on incomplete data and because they are rarely closely examined for consistency, they can lead to misconceptions.

The early days of physics education research (PER) saw great efforts expended to identify common misconceptions. In 1999, McDermott and Redish compiled a resource letter in which one section lists 115 references related to conceptual understanding and misconceptions in introductory university physics. For example, a very common misconception is to believe that when an object is tossed vertically upward, there is no force acting on it when it is at the apex of its trajectory, before it starts to travel back downward (Clement, 1982). The development of Discipline-Based Education Research (DBER) in other science and engineering fields followed a similar path, with examination of conceptual understanding and unearthing of common misconceptions being standard themes in the initial phases (see the National Academies review of DBER by Singer, Nielsen, & Schweingruber, 2012).

Misconceptions can be very difficult to uproot and replace with a robust understanding (Posner, Strike, Hewson, & Gertzog, 1982). Traditional teaching methods (lecturing) have been shown to have little impact on the development of conceptual understanding, even when students are successful at standard exams (see, for example, Hake, 1998). This realization has led to the development of variety of teaching tools and techniques that support active learning (e.g., Mazur, 1997; McDermott & Shaffer, 2002; Sokoloff & Thornton, 2004). In order to assess the effectiveness of various pedagogical practices at bringing about conceptual change, and thereby decide which to keep and which to discard, research-based and validated multiple-choice tests called concept inventories (CIs) have been developed in many disciplines and sub-disciplines in science and engineering. Examples of concept inventories include the Chemical Concepts Inventory (Mulford & Robinson, 2002), the Calculus Concept Inventory (Epstein, 2007), the Conceptual Inventory of Natural Selection (Anderson, Fisher, & Norman, 2002), and the Astronomy Diagnostic Test (Hufnagel, 2002) to name but a few. The reliability of a CI is an important issue for people who wish to use it as a diagnostic tool. It must accurately reflect the understanding of all students taking the test, and only their understanding of the material.

Arguably the best-known and most widely used CI in physics is the Force Concept Inventory (FCI), a diagnostic test designed to measure conceptual understanding of Newtonian mechanics, originally created by Hestenes, Wells, and Swackhammer (1992). In it, students are forced to choose between answers consistent with a Newtonian understanding of physics and answers that are consistent with common misconceptions about motion or mechanics. The FCI's ability to highlight specific student misconceptions has made this test a standard for measuring the efficacy of introductory physics teaching, both in the context of research on pedagogical approaches (see, for example, Beichner *et al.*, 2007; Brewe *et al.*, 2010; Caballero *et al.*, 2012; Crouch & Mazur, 2001; Cummings, Marx, Thornton, & Kuhl, 1999; Savinainen & Scott, 2002) and in action research by physics instructors seeking to monitor their teaching effectiveness (private communications from numerous physics instructors).

The validity and the reliability of the FCI have been established (Hake, 1998; Hestenes et al., 1992), though not without criticism (e.g., Heller & Huffman, 1995; Huffman & Heller, 1995). However, in recent years, questions have arisen concerning gender and the FCI. Madsen, McKagan, and Sayre (2013) describe the literature that exists comparing the performance of

male and female students on the FCI. The research suggests that there exists a significant gender gap between male and female results on the FCI, which may not exist in final class examination scores (Docktor & Heller, 2008). Madsen et al. (2013) also provide evidence from multiple studies that gaps exist between the gains of male and female students in FCI scores. This suggests that Force Concept Inventory results may not solely reflect the students' understanding of Newtonian mechanics, but also their gender.

Despite the consensus that the FCI shows bias towards male students (Madsen et al., 2013), a singular underlying cause has not been identified. The inventory's questions have been investigated using many methods, notably through differential item functioning (Dietz, Pearson, Semak, & Willis, 2012) and alternative question construction (McCullough & Meltzer, 2001). Both studies gave evidence that the FCI had questions that favoured one gender over the other, but neither were able to conclude firmly how students of a certain gender would be affected by a certain question. Kost, Pollock, & Finkelstein (2007), in contrast, found that the bias may not be the result of the FCI, but rather the result of different proficiencies in related areas like mathematics. The meta-analysis of the research on gender and concept inventories by Madsen et al. (2013) suggested that no one factor can be seen as dominant; instead, they suggest that the gender gap on the Force Concept Inventory is the accumulation of many small educational and psychological factors.

There are very few non-US data among the studies of the gender bias present in FCI scores. Bates et al. (2013) performed a notable study at three universities in the United Kingdom, which revealed a gender gap on pre-instruction FCI scores and FCI gains in favour of males in that selection of British universities. The gender gap also narrowed after instruction in the tested UK universities, a trend seen in some, but not all, US studies (Bates et al., 2013). The purpose of this study is to examine the Force Concept Inventory scores of male and female undergraduate students and examine the existence of a gender bias in the learning gains after undergoing an introductory physics course in a Canadian context. In addition the study aims to separate classes of different teaching styles and subject foci, to examine how these class elements affect the gender gap in these Canadian undergraduate FCI scores.

Method

This study has been reviewed and approved by the university Research Ethics Board and is on file as REB 2015-099.

Classes

The classes under study were all billed to students as introductory physics. Students were required to be concurrently enrolled in an introductory calculus course, or to have previously completed one. Engineering students were required to have completed high-school physics to grade 12, but no high-school physics was required of science students. Classes were held over three, fifty-minute sessions each week during a twelve-week semester. Classes prior to 2008 were separated by student discipline (science, engineering, non-science) and accompanied by a single, catch-all "enriched" class. Latterly, classes have been separated instead by interest, the course titles being "Introductory Physics I - Health & Life Science Interest," "Introductory Physics I - Physical Science Interest," and "Foundations of Physics for Engineers." Some or all of instructor-offered tutorials, office-hours, and peer-assisted learning programs were available

DOI: https://doi.org/10.5206/cjsotl-rcacea.2017.1.9

to students for physics practice outside of the classroom in the various classes, depending on the preference of the instructor.

Participants

The subjects of this study were students in thirteen first year physics courses over a nine-year period at a mid-sized, comprehensive, Canadian university. The majority, but not all, students were in their first year of study. While students were encouraged to take the accompanying lab course, it was not a requirement. The gender of participants was not self-reported on the test, but rather collected from the instructor or university records afterwards. Specific class information pertaining to teaching style and subject focus was also retained. In total each student's pre-instruction score, post-instruction score, gender, and class were recorded.

Procedure

The Force Concept Inventory was given to all students present in these courses during scheduled course hours. The test was first written by students early in the school term before formal instruction on the material, and written again at the end of the term. Students were informed that the test had no bearing on their course mark, and generally given 30 minutes to complete the assessment. Other than basic identification, there was no additional information requested alongside the FCI assessment. If a student did not write both FCI tests, or if they did not complete both tests (answering up to the last question), or if their gender was unknown, the student's marks were not included in the analysis.

Class Types

The research conducted not only aimed to study the gender disparity in learning gains on the FCI across multiple introductory physics classes, but also sought to see how different teaching styles and subject foci affected this disparity.

Teaching style. Madsen et al. (2013) found that interactive engagement methods showed potential for eliminating the gender gap on the Force Concept Inventory, but no conclusive pattern could be found. Interactive engagement methods in this study took the form of flipped classroom or lectures with peer instruction during class time. Of the thirteen classes examined during this study, eight were lecture-based courses while the other seven applied some method of interactive engagement. Marks from the interactive engagement classes were combined, and then compared to the amalgamated marks from lecture-based courses, allowing for the comparison of FCI scores, gain and gender disparity between students who were taught with different teaching styles.

Subject focus. From 2008 onwards, introductory physics courses included in this study were billed as either having a focus on health and life science applications or physical science applications. This focus was achieved through use of in-class examples, demonstrations and homework problems. Four "life-science" classes and five "physical-science" classes were examined, along with a single course that was a combined section of life science and physical science students. Comparisons were made between these three treatments to observe the effect of a life science subject focus on gender disparity. Engineering was not analysed as a separate

subject focus, because the only data available were for a single, small, separated section (although there were engineering students in several of the other classes).

Computational Tools

For the data analysis, a python 2.7 script (Pérez & Granger, 2007) was implemented through the Anaconda distribution (Continuum Analytics, 2015) using the modules:

- xlrd 0.9.4 and xlwt 1.0.0, for reading and writing the scores and results;
- scipy 0.15.1 and numpy 1.9.2, for their array manipulation and statistics

capabilities (Jones, Oliphant, Peterson, et al., 2001; Oliphant, 2007; van der Walt et al., 2011); and

- matplotlib 1.4.3 (Hunter, 2007) and seaborn 0.6.0 (Waskom, 2012), for creating histograms.

Quantitative Methods

Measured statistics. The following data and statistics were recorded and analysed in order to compare FCI performance between genders and gender performance between classes: (a) pre-instruction and post-instruction score distributions, (b) the distribution of student gains, and (c) the average gain of those distributions.

Learning gains. The primary statistic of interest in this study was the difference in learning gains between male and female students. Learning gains were measured through the normalised change between students' pre-instruction and post-instruction FCI scores, based on Marx and Cummings' (2007) refinement of the "Hake gain" (Hake, 1998):

$$c = \begin{cases} \frac{(\text{post-pre})}{(\text{1-pre})} & \text{post} > \text{pre} \\ \frac{(\text{post-pre})}{(\text{pre})} & \text{post} < \text{pre} \\ \text{drop} & \text{post} = \text{pre} = 1 \\ 0 & \text{post} = \text{pre} \neq 1 \end{cases} \qquad (1)$$

This is a slight philosophical adjustment of Marx and Cummings' original equation as students who had initial and final scores of zero would not be discarded. We note that there are no such students among our data, however.

Two measures of a class gain are possible: one can either take the normalised change of the average class marks or find the average of all the students' normalised changes. Here we use the average normalised change, which Marx and Cummings suggest is better at indicating class gain.

Score matching. Coletta and Phillips (2005) found that there was a significant, positive correlation between FCI gains and the pre-instruction score. In previous studies female students consistently score lower than male students on the initial FCI test (Bates et al., 2013; Dietz et al., 2012). To correct for this association, each set of male and female scores was matched, so that for each female student a male student with the same pre-instruction FCI score was found. This resulted in samples of the male and female students who shared identical pre-instruction score distributions. The comparison of male and female students was then repeated, with only matched male and female students counted. By comparing the results when counting the entire population

DOI: https://doi.org/10.5206/cjsotl-rcacea.2017.1.9

of tested students to this matched sample, the effect of dissimilar pre-instruction FCI scores could be seen. The same matching approach was used in the comparison of classes with different teaching style and in comparing classes with different subject foci.

Normalcy of data. Histograms of the data displayed distributions that had non-zero skew and kurtosis (Figure 1). Since the data were not normal, the use of non-parametric statistical testing was indicated. The tests used in this study were the Mann-Whitney *U*-Test and Cliff's delta test for dominance.

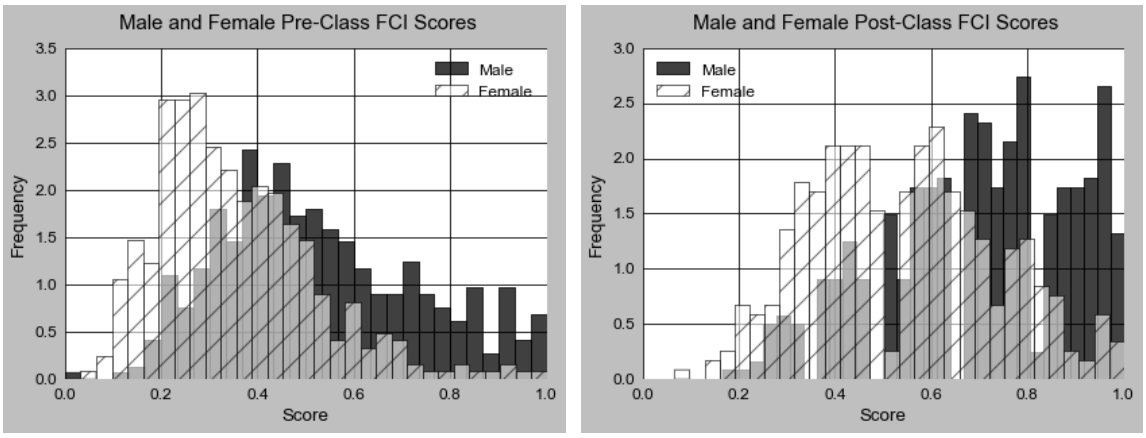

Figure 1. FCI scores of male and female students before and after instruction. The distributions were non-normal, preventing the use of parametric testing.

Statistical significance. Student score and gain distributions were used to compare the various gender and class-type groupings. Significance tests between gains and pre-class FCI scores were primarily accomplished using a Mann-Whitney *U*-test, which tests the null hypothesis that two distributions are the same (Mann & Whitney, 1947). The comparisons were performed in python with the scipy.stats.mannwhitneyu() function, which returned a *U*-value and a one-sided *p*-value. Significance testing allowed a meaningful way to compare the gain distributions between genders and classes, indicating whether a statistically significant difference existed between groups.

Effect size. Effect size measurements were used to give a quantifiable indication of how much difference was seen between groups. The most common measure of effect size, Cohen's *d*, is a parametric test (Peng, Chen, Chiang, & Chiang, 2013). Romano, Kromrey, Coraggio, Skowronek, & Devine (2006) suggest the use of Cliff's delta as an alternative, non-parametric measure of effect-size. Cliff's delta is a measure devised by Norman Cliff (1993) to quantify the "dominance" of one distribution over another, expressing how much overlap exists between the two data sets. The δ value for the dominance of distribution *a* compared to distribution *b* is calculated from the expression

$$\delta = \frac{\#(a_i > b_j) - \#(a_i < b_j)}{n_a n_b} \qquad (2)$$

where # denotes the 'number of times' and n_a and n_b refer to the sample sizes. (Cliff himself referred to this quantity as "the d statistic.") The magnitude of this value, which can range between 0 and 1, describes the degree of overlap between the two distributions; a value of one suggests that no overlap exists, indicating an extreme effect of the treatment. The sign captures

the direction of the effect, with positive values implying that the elements of distribution *a* are larger than the elements of distribution *b* (Macbeth, Razumiejczyk, & Ledesma, 2011).

Results

All Classes, Combined

Unmatched pre-instruction distributions. Table 1 shows that across the classes there exists an initial score disparity between male and female students which increases after instruction, as evidenced by the statistically significant ($p < 0.01$), higher normalised change for male students. The effect size measure ($\delta = 0.22$) suggests a small dominance of the male gain distribution.

Table 1
Male and Female FCI Learning Gains across Introductory Physics

Gender	*N*	Pre-Score	Av. Gain	*U*-Value	δ-Value
Male	434	0.51[a]	0.41[b]	63850*	0.22
Female	379	0.35[c]	0.30[d]		

Note. Pre-Score refers to the average pre-instruction score on the FCI. *U*, *p* and δ values calculated from comparing the gain distributions of male and female students, with the male distribution being held as distribution *a* (Eq. 2).
[a]$\sigma = 0.20$ [b]$\sigma = 0.29$ [c]$\sigma = 0.17$ [d]$\sigma = 0.25$
*$p < 0.01$.

Matched pre-instruction distributions. After matching the pre-instruction scores of male and female students, the dominance of the male gain distribution is reduced (to $\delta = 0.16$), as seen in Table 2; however the learning gains of female students remain significantly smaller than those of male students.

Table 2
Male and Female FCI Learning Gains across Introductory Physics; Matched Pre-Instruction FCI Distributions

Gender	*N*	Pre-Score	Av. Gain	*U*-Value	δ-Value
Male	255	0.41[a]	0.38[b]	27182.5*	0.16
Female	255	0.41[c]	0.31[d]		

[a]$\sigma = 0.16$ [b]$\sigma = 0.25$ [c]$\sigma = 0.16$ [d]$\sigma = 0.27$
*$p < 0.01$.

Teaching Style

Unmatched pre-instruction distributions. Table 3 shows a comparison of the genders within both lecture and interactive engagement classes, alongside a comparison of the two teaching styles' effects on each gender. A small effect appears showing higher gains for both male and female students in interactive engagement classes ($\delta_M = 0.23$, $\delta_F = 0.26$); however the dominance of the male gain distribution over the female gain distribution is slightly larger in the interactive engagement group ($\delta_{IE} = 0.26$) than in the lecture group ($\delta_L = 0.20$).

Table 3
Gender Differences in FCI Gains Between Teaching Styles

Class Type	Gender	*N*	Pre-Score	Av. Gain	*U*-Value	δ-Value
Interactive Engagement	Male	289	0.49[a]	0.44[b]	28995.5[*]	0.26
	Female	272	0.33[c]	0.33[d]		
Lecture	Male	133	0.54[e]	0.32[f]	5334[*]	0.20
	Female	100	0.35[g]	0.22[h]		
Comparison[i]	Male: Int/Lec	-	-	-	14684.5[*]	0.23
	Female: Int/Lec	-	-	-	10138.5[*]	0.26

[a]σ =0.20 [b]σ =0.30 [c]σ =0.19 [d]σ =0.28 [e]σ =0.21 [f]σ =0.27 [g]σ =0.16 [h]σ = 0.24

[i]Comparison of Interactive Engagement and Lecture gain distributions

[*]$p < 0.01$.

Matched pre-instruction distributions. When the pre-class FCI score distributions are matched, the average gains of female students increase (Table 4). Nevertheless, there is still a small dominance of male gains over female gains in both lecture and interactive engagement courses ($\delta_L = 0.23$ *vs.* $\delta_{IE} = 0.17$). As was the case for unmatched data, the gains are higher for both male and female students in the interactive engagement classes than in the lectures ($\delta_M = 0.29$, $\delta_F = 0.37$).

Table 4
Gender Differences in FCI Gains Between Teaching Styles; Matched Pre-Instruction FCI Distributions

Class Type	Gender	*N*	Pre-Score	Av. Gain	*U*-Value	δ-Value
Interactive Engagement	Male	173	0.39[a]	0.42[b]	12442.5[*]	0.17
	Female	173	0.39[c]	0.35[d]		
Lecture	Male	68	0. 45[e]	0.34[f]	1777[**]	0.23
	Female	68	0. 45[g]	0.23[h]		
Comparison[i]	Male: Int/Lec	124	0.53	0.48 /0.34	5489.5[*]	0.29
	Female: Int/Lec	92	0.35	0.36 /0.20	2681[*]	0.37

[a]σ =0.14 [b]σ =0.24 [c]σ =0.14 [d]σ =0.25 [e]σ =0. 17 [f]σ =0.27 [g]σ =0.17 [h]σ = 0.30

[i]Comparison of Interactive Engagement and Lecture gain distributions

[*]$p < 0.01$, [**]$p < 0.05$.

Subject Focus

Nine of the classes, for which data were collected, had subject-focused applications used in classroom teaching: four were physical science based, four were life science based, and one used a combination of both subjects in applications.

Unmatched pre-instruction distributions. For all three types of classes, male gains were significantly higher than female gains (Table 5). For the life science classes, the dominance of male gains ($\delta = 0.32$) was moderate, greater than for physical science classes ($\delta = 0.21$), and for the combination class ($\delta = 0.23$). When comparing the subject foci pairwise for both genders,

a significant different was only found between the gains of female students in the combined class compared to female students in the physical science classes ($\delta = 0.23$).

Table 5

Gender Differences in FCI Gains across Subject Foci

Class Type	Gender	*N*	Pre-Score	Av. Gain	*U*-Value	δ-Value
Health & Life Sciences	Male	112	0.47[a]	0.45[b]	5786.5[*]	0.32
	Female	152	0. 33[c]	0.31[d]		
Physical Sciences	Male	64	0.65[e]	0.40[f]	1062.5[**]	0.21
	Female	42	0.39[g]	0.26[h]		
Combination	Male	43	0. 54[i]	0.47[j]	1039[*]	0.23
	Female	63	0. 36[k]	0.36[l]		
Comparison[m]	Male: P/L	-	-	-	3238[***]	-0.10
	Female: P/L	-	-	-	2774.5[***]	-0.13
	Male: C/L	-	-	-	2251[***]	0.07
	Female: C/L	-	-	-	4193[***]	0.12
	Male: C/P	-	-	-	1197.5[***]	0.13
	Female: C/P	-	-	-	1020.5[**]	0.23

Note. P, L and C refer to the Physical Science, Health & Life Sciences, and Combination groups respectively.

[a]σ =0.19 [b]σ =0.26 [c]σ =0.17 [d]σ =0.24 [e]σ =0.21 [f]σ =0.34 [g]σ =0.22 [h]σ =0.27 [i]σ = 0.24, [j]σ = 0.31 [k]σ = 0.13 [l]σ = 0.30

[m]Pairwise comparison of the gains between subjects, with the first group considered distribution *a* (Eq. 2).

*$p < 0.01$. **$p < 0.05$. ***$p > 0.05$.

Matched pre-instruction distributions. The top three rows of Table 6 show the gain comparison between male and female students matched within each subject focus. The lower three rows compare matched students of a single gender across classes with different focus. Mann-Whitney significance testing found $p < 0.05$ only in the comparison between male and female students in the Health & Life Science class and between female students in the physical science and life science classes. Cliff's δ calculations showed moderate effects for female students when comparing the combined or life science courses to physical science courses, with the combined and life science offerings dominating. This domination of combined and life science against physical science gain distributions was also present in male students, but was smaller. All comparisons of subject foci are tempered by low *N*.

DOI: https://doi.org/10.5206/cjsotl-rcacea.2017.1.9

Table 6
Gender Differences in FCI Gains across Subject Foci; Matched Pre-Instruction FCI Distributions

Class Type	Gender	*N*	Pre-Score	Av. Gain	*U*-Value	δ-Value
Health & Life Sciences	Male	81	0.41[a]	0.45[b]	2682**	0.18
	Female	81	0.41[c]	0.37[d]		
Physical Sciences	Male	16	0.59[e]	0.51[f]	101.5***	0.21
	Female	16	0.59[g]	0.38[h]		
Combination	Male	25	0. 41[i]	0.39[j]	306***	0.02
	Female	25	0. 41[k]	0.37[l]		
Comparison[m]	Male: P/L	35	0.58	0.41 /0.49	527.5***	-0.14
	Female: P/L	32	0.31	0.24/ 0.39	338**	-0.34
	Male: C/L	10	0.52	0.52 /0.52	50***	0.0
	Female: C/L	25	0.34	0.40 /0.44	283.5***	-0.09
	Male: C/P	9	0.62	0.64 /0.46	28.5***	0.31
	Female: C/P	13	0.27	0.39 /0.25	48.5***	0.43

Note. P, L and C refer to the Physical Science, Health & Life Science, and Combination groups respectively.
[a]σ =0.15 [b]σ =0.24 [c]σ =0.15 [d]σ =0.27 [e]σ =0.23 [f]σ =0.32 [g]σ =0.23 [h]σ =0.28 [i]σ = 0.17 [j]σ = 0.32 [k]σ = 0.17 [l]σ = 0.27
[m]Comparison of the gains between subjects, with the first group considered distribution *a*.
*p < 0.01. **p < 0.05. ***p > 0.05.

Discussion

All Classes, Combined

There is a significant, but small, dominance of male gains over female gains on the Force Concept Inventory. The *U*-value assesses the significance of the difference in distributions; the low *p*-value allows for the confident assertion that male and female gain distributions were different. The δ value quantifies this difference: δ = 0.22 (Table 1).

To eliminate the effect of male students entering the physics classroom with higher FCI scores (Kost et al., 2007), matching was used to compare gender samples with identical pre-instruction score distributions. The resulting reduction in gain disparity suggests that the slightly higher concentration of high-proficiency students in the male group had some effect on the comparison, but the change in δ of 0.06 (Tables 1, 2) indicates this effect is slight. The matched scores confirm the initial result: male students show greater improvements in FCI scores (greater gains) after a term of instruction than their female counterparts.

The disparity in gain seen between male and female students is consistent with prior research at US universities. Madsen et al. (2013) suggest that across the literature there was an average gain of 0.43 for men and 0.37 for women when considering nine interactive engagement courses in the study and two lecture based courses (ratio of interactively engaged students to

lectured students 7.4:1). In this study men and women had an average gain of 0.38 and 0.31 respectively, when correcting for initial proficiency (Table 2). The ratio of students in interactive-engagement classes (8) to lecture-based classes (7) in this research was 2.4:1. Considering the results in Tables 3 and 4 suggests that the larger proportion of lecture-based courses could be the source of the slightly lower gains when compared to previous research.

Complete assessment data for the single combined class (235 students) were also checked for gender bias. Distributions of marks in that class were statistically indistinguishable for male and female students (a) in the term tests (p = 0.20, δ = 0.10), which mostly consisted of conceptual questions like the FCI, (b) in the concept-testing section of the final exam (p = 0.57, δ = 0.06), (c) in the calculational part of the final exam (p = 0.66, δ = 0.04), and (d) in the overall mark for the course (p = 0.64, δ = 0.04). Although complete assessment data have not been analysed for all classes, this check implies that the FCI is a different assessment tool for men and women, rather than there being some systemic gender bias at the institution in this study.

Overall, the results suggest that the Force Concept Inventory has similar biases towards males, in this Canadian university, to those that have been demonstrated in previous studies from the US and the UK. This is important as the Force Concept Inventory has been widely used outside of its native US, but has not been widely tested in other countries. Further enquiry into the specific action of the Force Concept Inventory on Canadian students is warranted to understand whether the test truly reflects their physics knowledge.

Teaching Style

The increase in Force Concept Inventory gains caused by the use of interactive engagement methods in the classroom has been widely studied (Hake, 1998; Madsen et al., 2013; see also McDermott & Redish, 1999); the results of our investigation are no surprise in that regard.

Previous literature has drawn mixed conclusions on the effectiveness of adding interactive engagement methods at reducing the gender gap. Lorenzo, Crouch, and Mazur (2006) found a large reduction in gender gaps in interactive engagement courses; in contrast Pollock, Finkelstein, and Kost (2007) found no such reduction and even saw instances of increased gender disparity in gains. Before matching the scores, the interactive engagement courses appeared to have a larger gender gap than the lecture based courses, with a small increase in gain being demonstrated by both genders (Table 3). Matching the pre-instruction FCI score distributions, however, reversed the situation for the teaching-style comparison, while the improvements in inter-class gain associated with interactive engagement methods for both genders rose to moderate levels (Table 4). The matched scores allow for a clear comparison of both gender and teaching style effects, and suggest that interactive course elements improve learning gains, but do not eliminate the gender gap in the reported context.

Subject Focus

The gender gap in gains within the physical science and combined group was small, with a magnitude similar to those seen in the other analyses, of $\delta \approx 0.2$. The life science class, however, had a moderate gender gap between male and female students, which was higher than all of the other effect sizes between genders (Table 5). When the pre-instruction scores are matched (Table 6), the numbers in each group are smaller and additional caution is warranted

DOI: https://doi.org/10.5206/cjsotl-rcacea.2017.1.9

when interpreting the results; the gender gap in the life science class is the only one that remains significant. Comparisons between classes hint that female students have smallest gains in physical science classes (compared to the combined class in Table 5 or the life science class in Table 6, where the pre-instruction scores are matched).

Interpreting the results from Tables 5 and 6 is difficult due to a variety of confounding factors. A major issue was the absence of physical science courses that used interactive engagement and life science courses that were lecture based. Four teachers' classes were represented in the study; two of these teachers taught physical science classes and two of them taught life science classes. This leaves the possibility that the specific teaching style of the teacher, including their use of lecturing or interactive engagement, confuses any comparison between subject foci. The combined course is affected by these factors in a large way; while both physical science and life science students each came from four classes held in different years, the combined course was a single, one-time offering taught by a single teacher using a flipped classroom approach. While these factors do not impact the ability to discuss the gender gap within each class type, they make it difficult to interpret the comparisons between the subjects.

Classroom size and demographics also differed dramatically between the different offerings of physics. Life-science interest courses had large classes with a slight majority of female students; in contrast physical-science courses had smaller classes with an overwhelming male majority. The life-science courses found a moderately sized gap between the male and female gain distributions, while the physical-science courses had a small gap. Madsen et al. (2013) find no correlation between classroom demographics and the reduction of the gain gender gap, suggesting that gaps in learning gains may not be solved with a simple change of classroom demographics. The moderate size of the life-science gender gap (Table 5) supports this; having more female students than male students alone does not reduce the gender gap.

The differences between the gender gap in life-science gains and the gender gap in physical science or combined gains may be linked to a complex interaction of many factors. The possibility exists that student perception of the class types as separated based on factors beyond the stated subject focus may have led to self-selection and stereotyping. For example, the course billing 'physical-science interest' may have subtly implied 'for physicists'. Since female students may already associate physics as masculine and negative (Kessels, Rau, & Hannover, 2006), this could lead to students being dissuaded from the physical science offering. This effect has the potential to create a self-fulfilling prophecy, as students internalise the "non-physicist" class as having a majority of female students. This may, in turn, reinforce the stereotype that women are unwelcome in physics. This offers an explanation of the moderate gap between female and male student gains in the life-science courses, as stereotype effects have been shown to have a significant impact on the scores of students in marginalised or under-represented groups (Shapiro & Williams, 2012; Steele, 1997). There is a need for future qualitative research to understand whether splitting students by subject interest leads to unintended connections being drawn between classroom demographics and suitability to the study of physics.

Conclusion

Overall, there is a gap on the Force Concept Inventory between male and female introductory physics students when entering the classroom at a medium-sized, Canadian university. Male students also appear to have a small advantage in learning gains, leading to an increase in the gender gap on post-instruction FCI tests. This effect is marginally reduced when

examining male and female students with identical starting Force Concept Inventory scores, but remains visible.

Further examination of the gender gap, by splitting the classes into lecture based or interactive engagement courses, shows that the small dominance of male over female gains occurs across teaching styles. The gender gap remains, despite the use of interactive engagement, when considering students with identical pre-instruction scores. This mirrors a similar FCI study by Pollock et al. (2007), in which gender gaps in the classroom were not reduced by interactive engagement course elements, and contrasts the findings of Lorenzo et al. (2006), in which gaps were reduced by adding engagement elements. Interactive engagement courses, however, did provide a moderate gain increase for both genders compared to their lecture counterparts, consistent with previous studies on the effect of interactive engagement course elements on FCI gains (Hake, 1998; Madsen et al., 2013).

Comparing classes based on the subject focus of physics applications reveals the greatest gap between male and female gain in life-science classes. This was the only gap to remain significant in matched samples. These results are not necessarily indicative of the effect of the addition of life-science applications into instruction, as teaching style, classroom demographics and course stereotypes potentially act as confounding variables.

Given that there was no apparent gender bias in the in-course assessments (insofar as the data were available), future work will concentrate upon the format of the FCI itself and on its implementation, in order to unearth the source of the puzzling gender disparity. Our data provide a cautionary tale for any who use concept inventories in their research or as part of their self-assessment in teaching. We encourage careful consideration of whether gender or other social factors may be impacting the results.

References

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the Conceptual Inventory of Natural Selection. *Journal of Research in Science Teaching*, *39*(10), 952–-78. https://doi.org/10.1002/tea.10053

Bates, S., Donnelly, R., MacPhee, C., Sands, D., Birch, M., & Walet, N. R. (2013). Gender differences in conceptual understanding of Newtonian mechanics: A UK cross-institution comparison. *European Journal of Physics*, *34*(2), 421. https://doi.org/10.1088/0143-0807/34/2/421

Beichner, R. J., Saul, J. M., Abbott, D. S., Morse, J. J., Deardorff, D. L., Allain, R. J.., Bonham, S. W., Dancy, M. H., & Risley, J. S. (2007). The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project. In *Research-Based Reform of University Physics* (1). Retrieved from www.compadre.org/Repository/document/ServeFile.cfm?ID=4517&DocID=183

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). How people learn: Brain, mind, experience, and school, expanded edition. Washington, DC: The National Academies Press.

Brewe, E., Sawtelle, V., Kramer, L., O'Brien, G., Rodriguez, I., & Pamelá, P. (2010). Toward equity through participation in Modeling Instruction in introductory university physics. *Physical Review Physics Education Research*, *6*(1), 010106. https://doi.org/10.1103/PhysRevSTPER.6.010106

Caballero, M. D., Greco, E. F., Murray, E. R., Bujak, K. R., Marr, M. J., Catrambone, R., Kohlmyer, M. A., & Schatz, M. F. (2012). Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study. *American Journal of Physics*, *80*(7), 638-644. https://doi.org/10.1119/1.12989

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics*, *50*(1), 66-70. https://doi.org/10.1119/1.12989

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494. https://doi.org/10.1037/0033-2909.114.3.494

Coletta, V. P., & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics*, *73*(12), 1172-1182. https://doi.org/10.1119/1.2117109

Continuum Analytics. (2015, July 2). *Anaconda 2.3.0* (Python 2.7.10 ed.) [Computer software]. Retrieved from http://continuum.io/downloads

Crouch, C. H., & Mazur, E. (2001) Peer instruction: Ten years of experience and results. *American Journal of Physics*, *69*, 970-977. https://doi.org/10.1119/1.1374249

Cummings, K., Marx, J., Thornton, R., & Kuhl, D. (1999) Evaluating innovation in studio physics. *American Journal of Physics*, *67*, S38-S44. https://doi.org/10.1119/1.19078

Dietz, R. D., Pearson, R. H., Semak, M. R., & Willis, C. W. (2012). Gender bias in the force concept inventory? *AIP Conference Proceedings*, *1413*(1), 171-174. https://doi.org/10.1063/1.3680022

Docktor, J., & Heller, K. (2008). Gender differences in both Force Concept Inventory and Introductory Physics Performance. In *American Institute of Physics Conference Series, 1064*, 15-18. https://doi.org/10.1063/1.3021243

Epstein, J. (2007). Development and validation of the Calculus Concept Inventory. *Proceedings of the Ninth International Conference on Mathematics Education in a Global Community*, pp. 165-170.

Hake, R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*, 64-74. https://doi.org/10.1119/1.18809

Heller, P., & Huffman, D. (1995). Interpreting the Force Concept Inventory: A reply to Hestenes and Halloun. *The Physics Teacher*, *33*, 503-511. https://doi.org/10.1119/1.2344279

Hestenes, D., Wells, M., & Swackhammer, G. (1992). Force Concept Inventory. *The Physics Teacher*, *30*(3), 141-158. https://doi.org/10.1119/1.2343497

Huffman, D., & Heller, P. (1995). What does the Force Concept Inventory actually measure? *The Physics Teacher*, *33*, 138-143. https://doi.org/10.1119/1.2344171

Hufnagel, B. (2002). Development of the astronomy diagnostic test. *Astronomy Education Review*, *1*(1), 47-51. https://doi.org/10.3847/AER2001004

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, *9*(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

Jones, E., Oliphant, T., Peterson, P. and others (2001). *SciPy: Open source scientific tools for Python*. Retrieved from http://www.scipy.org/

Kessels, U., Rau, M., & Hannover, B. (2006). What goes well with physics? Measuring and altering the image of science. *British Journal of Educational Psychology*, *76*(4), 761-780.

Kost, L. E., Pollock, S. J., & Finkelstein, N. D. (2007). Investigating the source of the gender gap in introductory physics. In *2007 Physics Education Research Conference*, 951, 136-139. https://doi.org/10.1348/000709905X59961

Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, *74*(2), 118-122. https://doi.org/10.1119/1.2162549

Macbeth, G., Razumiejczyk, E., & Ledesma, R. D. (2011). Cliff's delta calculator: A non-parametric effect size program for two groups of observations. *Universitas Psychologica*, *10*(2), 545-555.

Madsen, A., McKagan, S. B., & Sayre, E. C. (2013). Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Physical Review Special Topics Physics Education Research*, *9*, 020121. https://doi.org/10.1103/PhysRevSTPER.9.020121

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50-60. https://doi.org/10.1214/aoms/1177730491

Marx, J. D., & Cummings, K. (2007). Normalized change. *American Journal of Physics*, *75*(1), 87-91. https://doi.org/10.1119/1.2372468

Mazur, E. (1997). *Peer instruction: A user's manual*. Upper Saddle River, NJ: Prentice-Hall.

McCullough, L., & Meltzer, D. (2001). Differences in male/female response patterns on alternative-format versions of the force concept inventory. In *Proceedings of the 2001 Physics Education Research Conference*, 103-106. https://doi.org/10.1119/perc.2001.pr.013

McDermott, L. C., & Redish, E. F. (1999). Resource letter: PER-1: Physics education research. *American Journal of Physics*, *67*, 755-767. https://doi.org/10.1119/1.19122

McDermott, L. C., & Shaffer, P. S. (2002) *Tutorials in introductory physics.* Prentice-Hall, NJ, USA.

Mulford, D. R., Robinson, W. R. (2002). An inventory for alternate conceptions among first-semester general chemistry students. *Journal of Chemical Education*, *79*(6), 739-744. https://doi.org/10.1021/ed079p739

Oliphant, T. (2007). Python for scientific computing. *Computing in Science and Engineering*, *9*, 10-20. https://doi.org/10.1109/MCSE.2007.58

Peng, C.-Y. J., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The Impact of APA and AERA Guidelines on Effect Size Reporting. *Educational Psychology Review*, *25*(2), 157-209. https://doi.org/10.1007/s10648-013-9218-2

Pérez, F., & Granger, B. E. (2007). IPython: A system for interactive scientific computing. *Computing in Science and Engineering*, *9*, 21-29. https://doi.org/10.1109/MCSE.2007.53

Pollock, S. J., Finkelstein, N. D., & Kost, L. E. (2007). Reducing the gender gap in the physics classroom: How sufficient is interactive engagement? *Physical Review Special Topics Physics Education Research*, *3*, 010107. https://doi.org/10.1103/PhysRevSTPER.3.010107

Posner, G. J., Strike, K. A., Hewson, P. W., Gertzog, W. A. (1982). Accomodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211-227. https://doi.org/10.1002/sce.3730660207

Romano, J., Kromrey, J. D., Coraggio, J., Skowronek, J., & Devine, L. (2006, October). Exploring methods for evaluating group differences on the NSSE and other surveys: Are the t-test and Cohen's d indices the most appropriate choices? Paper presented at the Annual Meeting of the Southern Association for Institutional Research, Arlington, Virginia.

Savinainen, A, & Scott, P. (2002) Using the Force Concept Inventory to monitor student learning and to plan teaching. *Physics Education*, *37*(1), 53-58. https://doi.org/10.1088/0031-9120/37/1/307

Shapiro, J. R., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles*, *66*(3-4), 175-183. https://doi.org/10.1007/s11199-011-0051-0

Singer, S. R., Nielsen, N. R., & Schweingruber, H. A. (2012). Discipline-based education research: Understanding and improving learning in undergraduate science and engineering. Washington, DC: The National Academies Press, 282.

Sokoloff, D. R., & Thornton, R. K. (2004). *Interactive learning demonstrations: Active learning in introductory physics*. Hoboken, NJ: John Wiley and Sons, 374.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613-629. https://doi.org/10.1037/0003-066X.52.6.613

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, *13*, 22-30. https://doi.org/10.1109/MCSE.2011.37

Waskom, M. (2012). Seaborn: Statistical data visualization. Retrieved from http://stanford.edu/mwaskom/software/seaborn/index.html